Solution to Midterm Question on Softmax Backpropagation

March 7, 2020

Recall that the softmax function takes in a vector $(z_1, \ldots, z_D)$ and returns a vector $(y_1, \ldots, y_D)$. We can express it with the following equations, illustrated in the network shown below for $D = 2$.

$$\begin{aligned} r &= \sum_j \exp(z_j) \\ y_i &= \exp(z_i)/r \end{aligned}$$

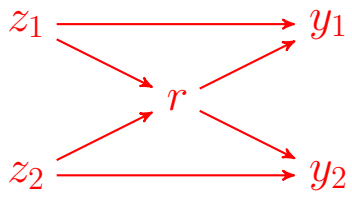

Figure 1: Softmax Computation Graph

The error is a function of the output nodes; we may write it as $E(y_1, y_2)$. We can assume that we have already computed the error derivatives with respect to the output nodes, as in the backpropagation algorithm:

$$\text{Assume } \frac{\partial E}{\partial y_1} \text{ and } \frac{\partial E}{\partial y_2}.$$

To find the gradient of the error function with respect to network weights that feed into the z_i variables, we need to find the gradient with respect to the z_i variables. This can be done using the multi-variate chain rule as follows. (See also the solution in the posted sample midterm.)

$$
\begin{aligned}
\frac{\partial y_i}{\partial r} &= -\exp(z_i)/r^2 \\
\frac{\partial E}{\partial r} &= \frac{\partial E}{\partial y_1} \cdot \frac{\partial y_1}{\partial r} + \frac{\partial E}{\partial y_2} \cdot \frac{\partial y_2}{\partial r} \\
&= \frac{\partial E}{\partial y_1} \cdot (-\exp(z_1)/r^2) + \frac{\partial E}{\partial y_2} \cdot (-\exp(z_2)/r^2)) \\
&= -\frac{1}{r^2}(\frac{\partial E}{\partial y_1} \cdot \exp(z_1 + \frac{\partial E}{\partial y_2} \cdot \exp(z_2)) \\
\frac{\partial E}{\partial z_1} &= \frac{\partial E}{\partial y_1} \cdot \frac{\partial y_1}{\partial z_1} + \frac{\partial E}{\partial r} \cdot \frac{\partial r}{\partial z_1} \\
&= \frac{\partial E}{\partial y_1} \cdot \exp(z_1)/r + \frac{\partial E}{\partial r} \cdot \exp(z_1) \\
\frac{\partial E}{\partial z_2} &= \frac{\partial E}{\partial y_2} \cdot \frac{\partial y_2}{\partial z_2} + \frac{\partial E}{\partial r} \cdot \frac{\partial r}{\partial z_2} \\
&= \frac{\partial E}{\partial y_2} \cdot \exp(z_2)/r + \frac{\partial E}{\partial r} \cdot \exp(z_2)
\end{aligned}
$$

I tried to make this problem easier and therefore suggested an approach where we think of the computation graph as a neural network with activation functions g. In that case we can apply the standard backpropagation formula. The problem is that to make the computation correct, we end up feeding to y_i the product of inputs rather than the sum as required by backpropagation. As we discussed in class, several of you followed this approach. We'll give bonus points for this. Come see me if you want me to explain it in detail.